Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 5, 908–923

0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.5.908

Adaptive Categorization in Unsupervised Learning

John P. Clapper
Humboldt State University

Gordon H. Bower
Stanford University

In 3 experiments, the authors provide evidence for a distinct category-invention process in unsupervised (discovery) learning and set forth a method for observing and investigating that process. In the 1st 2 experiments, the sequencing of unlabeled training instances strongly affected participants' ability to discover patterns (categories) across those instances. In the 3rd experiment, providing diagnostic labels helped participants discover categories and improved learning even for instance sequences that were unlearnable in the earlier experiments. These results are incompatible with models that assume that people learn by incrementally tracking correlations between individual features; instead, they suggest that learners in this study used expectation failure as a trigger to invent distinct categories to represent patterns in the stimuli. The results are explained in terms of J. R. Anderson's (1990, 1991) rational model of categorization, and extensions of this analysis for real-world learning are discussed.

The ability to discover patterns and regularities across multiple experiences is a fundamental component of human cognition. Often, people must acquire such generalizations through untutored observation, without external guidance or corrective feedback, a situation referred to as *unsupervised learning*. In contrast to standard experiments on supervised concept learning (e.g., Bruner, Goodnow & Austin, 1956; Smith & Medin 1981), in experiments on unsupervised learning there is no teacher or environmental agent to provide trial-by-trial feedback regarding the correct classification of the training stimuli. Rather, learners must ascertain, induce, or discover for themselves whatever predictive structure may exist within a given stimulus domain. This "discovery problem" is arguably the most interesting aspect of unsupervised learning and is the most important sense in which unsupervised learning differs from supervised learning.

We have two goals, one methodological and one theoretical, for the present article. First, we introduce a new experimental procedure that allows us to investigate unsupervised learning in an unobtrusive manner as it occurs over time. This method contrasts with others that merely assess the knowledge that participants have at one point in time, for example, following a prior training series. Second, we use this new procedure to gather data that distinguish among theories of unsupervised learning. Specifically, we present evidence that people learn about the predictive structure of our stimulus sets by inventing discrete categories rather than by gradually strengthening associations between (or learning rules about) co-occurring elements of the training instances.

John P. Clapper, Department of Psychology, Humboldt State University; Gordon H. Bower, Department of Psychology, Stanford University.

This research was supported in part by U.S. Air Force Office of Scientific Research Grant AFOSR-87-0282 and by National Institute of Mental Health Grant 1R37-47575 to Gordon H. Bower.

We thank Dorrit Billman, Thomas Wallsten, and three anonymous reviewers for comments on an earlier version of this article. We also thank Terry Nellis, Katherine Longueville, and Claudia Cole for their help in testing the participants in these experiments.

Correspondence concerning this article should be addressed to John P. Clapper, who is now at the Department of Psychology, California State University, 5500 University Parkway, San Bernardino, California 92407. E-mail: jclapper@csusb.edu

Theories of Unsupervised Learning

The task for any model of unsupervised learning is to examine a series of (unlabeled) training instances and generate a description (e.g., a set of rules or categories) that captures any patterns or regularities exhibited across those instances. For example, the learner might be presented with a series of stimuli designed according to the abstract specifications shown in Table 1. Here, *predictive structure* refers to correlations among the values of some attributes of the training stimuli (attributes 2, 4, 5, 7, 8, and 9 have correlated values and attributes 1, 3, and 6 have uncorrelated values).

People might learn these attribute correlations by incrementally tracking attribute-value covariations as successive instances are encountered. This approach would capture correlational patterns directly, in the form of rules or associations summarizing how often different attribute values tend to occur together across instances. By keeping and updating a memory record of which attribute values frequently co-occur with others, a correlation tracker would gradually accumulate information about the predictive (correlational) structure within a given stimulus domain. Examples of correlation trackers are the auto-association models of J. A. Anderson (1977; J. A. Anderson, Silverstein, Ritz, & Jones, 1977) and McClelland and Rumelhart (1985; Rumelhart, Hinton, & McClelland, 1986); the exemplar-storage models[1] of Hintzman

[1] Exemplar-storage models, of course, store whole instances rather than explicit feature correlations. However, in unsupervised learning tasks these models could compute ad hoc correlations as needed. For example, given some features of an instance, the model could fill in missing values by first retrieving stored instances that match the target instance on its known

Table 1
Sample Stimulus Set Illustrating Predictive (Correlational) Structure

		Attributes								
Pattern	Instance	1	2	3	4	5	6	7	8	9
A	1	2	1	2	1	1	2	1	1	1
	2	1	1	2	1	1	1	1	1	1
	3	1	1	1	1	1	2	1	1	1
	4	2	1	1	1	1	1	1	1	1
B	1	1	2	1	2	2	1	2	2	2
	2	2	2	2	2	2	1	2	2	2
	3	1	2	2	2	2	2	2	2	2
	4	2	2	1	2	2	2	2	2	2

Note. The rows 1 through 4 denote instances composed of 9 binary-valued attributes; the values of these attributes are denoted by "1" and "2."

(1986) and Medin and Shaffer (1978); and the rule (production)-based systems of Billman and Heit (1988), Davis (1985), and Zeaman and House (1963).

A second way that people might capture predictive structure would be based on first partitioning or segregating different subsets (categories) of instances within a given stimulus set and then accumulating information about consistencies within each subset or category (e.g., J. R. Anderson, 1990, 1991; Clapper & Bower, 1991; Fisher & Langley, 1990; Gluck & Corter, 1985). Depending on the theory, category representations might take the form of ideal instances or prototypes (e.g., Homa & Cultice, 1984), statistical estimates of central tendency and variability of each attribute in each category (e.g., Fried & Holyoak, 1984), attribute-value probabilities within each subset (J. R. Anderson, 1990, 1991; Clapper & Bower, 1991), or connection strengths between features and categories in a connectionist network (e.g., Rumelhart & Zipser, 1986). In such models, the basic principle involves the use of some event or metric for splitting off separate categories containing contrasting subsets of instances and then learning about the consistent structure within each subset.

Measures of Unsupervised Learning

A first step in distinguishing among theories of unsupervised learning is deciding on appropriate learning tasks and performance indices. In general, the more participants are told in advance about the structure of a given stimulus domain (e.g., that the stimuli can be sorted into a specific number of distinct categories based on correlational patterns), the less investigators can learn about how participants might have discovered such predictive structure on their own (the "discovery problem" referred to above). Furthermore, many of the rules and/or categories that people acquire from everyday experience are probably picked up incidentally, as a side effect of their interaction with their immediate task environment rather than as the result of a deliberate reasoning process. This suggests that an ideal unsupervised learning task should allow investigators the option of not informing participants about the predictive structure they are expected to learn; moreover, it should allow learning to be observed unobtrusively, without participants necessarily being aware that they are expected to discover any rules or categories within the stimulus set.

Standard methods of investigating unsupervised learning often force participants to search explicitly for predictive structure in the stimulus set and may even specify what that structure should look like, greatly reducing how much these methods can elucidate the discovery aspect of unsupervised learning. For example, this criticism applies to the continuous sorting or category construction tasks that are commonly used in studies of unsupervised learning (e.g., Fried & Holyoak, 1984). Another method for inferring what individuals have learned is to ask participants to estimate the covariation between features of the stimulus set (e.g., Kaplan & Murphy, 1999) or to judge which of a pair of test stimuli best matches the training stimuli they have seen up to that point (e.g., Billman & Knutson, 1996). If interspersed with the training instances, such tests obviously suggest to participants that they should be looking for covariations among features on later study trials. Thus, such methods are not very useful for tracking unsupervised learning in an "uncontaminated" manner as it occurs over multiple experiences.

Rather than assess learning continuously from the start of training, most researchers on unsupervised learning have separated their experiments into distinct training and transfer-testing phases; sorting or covariation judgment tasks are then introduced only during the transfer phase to assess earlier learning (e.g., Billman & Knutson, 1996; Kaplan & Murphy, 1999). The advantage of this method is that participants can be shown an initial training set under conditions that do not force them to search for co-occurrences or a prespecified number of categories. The major disadvantage of such transfer measures is that at best they allow the assessment of learners' knowledge only after it has been acquired, not as it develops throughout training. The inability to observe unsupervised learning throughout training is a serious limitation because different learning algorithms often differ radically in their predictions about the shape of the acquisition function, how long learning should take under different circumstances, how the training instances are processed as a result of current learning, and so on.

In earlier work (Clapper & Bower, 1991, 1994), we introduced a method that proved successful for tracking unsupervised learning in an unobtrusive manner without conveying to participants that there was any stimulus structure to be learned. In this procedure, referred to simply as *attribute listing*, participants are presented with complex stimulus patterns (e.g., pictures of novel insects that varied in several attributes) one at a time. The features of these stimuli were designed according to specifications similar to those in Table 1; most important, some attribute values of the stimuli

values and then predicting the value of the target (missing) attribute(s) that occurs most often in this set of matching instances (e.g., Hintzman's, 1986, MINERVA model). Such models would use memory essentially to compute correlations between the known attribute values of the current instance and each possible value of its unknown attributes and then to fill in the value with the highest probability. Thus, for present purposes exemplar-storage models can be grouped with correlation-tracking models, with the caveat that the correlations are not stored directly but rather are computed on demand from stored instances. Of course, this assumes that the models in question lack any ability to create new categories in response to novel or surprising instances and that they capture predictive structure within a domain by simply storing raw instances and computing feature correlations as needed.

covaried perfectly with other values, such that the stimuli were naturally divisible into examples of Pattern A versus Pattern B based on these correlations. As each insect instance (rows of Table 1) appeared, participants were asked to list (i.e., to write down) a few attributes of each insect that they judged would be most helpful for distinguishing it from the other insects they had seen in the series. The instructions were framed so as to motivate participants to list only those features that they considered highly informative for a later recognition-memory task (never actually given), while omitting uninformative features.

Our participants showed strong evidence of learning in this task. Individuals came increasingly to list the unpredictable (uncorrelated) attributes of the insect instances (these are the values of attributes 1, 3, and 6 in Table 1). Concomitantly, they showed progressively less listing of the correlated, highly predictable attributes of the insects (attributes 2, 4, 5, 7, 8, and 9 in Table 1). They came to display nearly optimal listing behavior: If the insects are characterized by patterns of intercorrelated attributes with a few varying unpredictable features, then to remember any given insect it suffices to list any one of the intercorrelated attributes plus all of the unpredictable attributes that identify this instance.

Testing Theories of Unsupervised Learning

In our earlier attribute-listing studies, we found a profound effect of the sequence or order in which instances of two patterns were presented to participants (Clapper & Bower, 1994). Unsupervised learning was far easier if participants saw a long block of instances all of one type before they ever saw instances of a second type. For example, Patterns A and B (from Table 1) would be learned much more readily if training instances were presented in a series such as *AAAAAAAABBBBBBBB* than if they were shown in a series such as *ABBABAABBABABAAB* (a similar result was reported by Medin & Bettger, 1994). It is important to note that learning was affected more strongly by the specific order in which training instances from different patterns were shown than by the sheer number of instances shown from a given pattern. Thus, learning of Pattern B was far greater in a series such as *AAAAAAAAABABBAB* than in a series such as *ABBABAABABABBAB*, in which the underlined *A* instances from the first sequence were replaced with *B*s. The first series produced much better learning of Pattern B despite the fact that more *B* instances were shown in the second series (Pattern A was also learned better in the first series).

Such category-sequence effects pose severe difficulties for unsupervised learning algorithms based purely on correlation tracking. Most damaging, these models update correlational measures (rules or associations) with each positive training instance and so expect the strength of a given pattern in memory to increase as a direct function of exposure to such instances. Such models appear unable to accommodate our results in which presenting additional instances of a pattern actually reduced learning of that pattern, as in the last example above.

By contrast, the category-invention approach provides a natural explanation for our sequence effects (Clapper & Bower, 1994). Assuming that presenting several *A*s in succession permits learning of their shared features, we theorize that participants would view the first *B* instance as very novel, as breaking the expected (prior) feature co-occurrences and variations. By hypothesis, this failed expectation, when of sufficient magnitude, serves as a trigger for the learner to create a new category (for the *B*s), separate from the category set up earlier to characterize the prior instances (of Pattern A). The intuition is that a surprising change in the type of instances encountered causes learners to generate a new category and to begin learning about its separate feature distributions. Without such a distinct, surprising event to signal the contrast between different subsets of instances, the learner might simply assign them all to a single group or category. Such an aggregated grouping would obscure the correlational patterns exemplified within the two classes of stimuli.

A visual illustration of this kind of category learning process is shown in Figure 1. In the top panel (Figure 1A), presenting a training instance from Pattern A causes the learner to create an initial category around that instance. After only a single example of *A*, this region is broad enough so that the first example of Pattern B, shown on Trial 2, falls within it and is accepted as a member of the same (A) category. This first category then remains sufficiently wide over subsequent trials—perhaps even expanding in size—to accommodate further examples of both *A* and *B*. This results in a single overgeneralized category that fails to capture the correlated attribute values of the *A*s and *B*s.

By contrast, the lower panel (Figure 1B) shows a case in which a large number of *A* instances are presented prior to the first *B*, causing the boundaries of the *A* category to shrink around the consistent dimensions of these instances. The first example of Pattern B violates these narrowed expectations and so falls outside the *A* category. Because of this major mismatch, a new category would be generated for that unusual *B* stimulus. Further examples of Pattern B are then assigned to this *B* category and further examples of Pattern A continue to be assigned to the *A* category.

The Study-Time Procedure

As noted, the attribute-listing procedure is unobtrusive and allows learning to be assessed continuously, so that distinct learning curves can be plotted for each correlational pattern in a given stimulus set. However, the number of attributes listed is under the learners' control, so we cannot conclude that an attribute that is not listed was not attended to or rehearsed. Moreover, attribute listing provides no information about how learners' performance capabilities (e.g., memory) have been altered because of their unsupervised learning. Also, attribute listing is a relatively novel procedure in concept-learning studies, suggesting a need to bolster our conclusions with converging evidence from other procedures.

For such reasons, we developed a second procedure for measuring unsupervised learning in an unobtrusive manner. In this new method, participants are shown a series of training instances (verbal feature lists rather than pictorial stimuli) and are asked to study each list for an immediate recognition-memory test. The stimuli are constructed according to specifications similar to those shown in Table 1, that is, distinct patterns (categories) are defined by several correlated attributes whereas other attributes vary independently across instances. The procedure allowed us to observe how much time people spend looking at (i.e., studying) each listed attribute value, as well as examining recognition accuracy for each item in the memory test that follows immediately. If participants learn the correlational patterns, then their behavior should exhibit two characteristics: First, participants should gradually reduce the

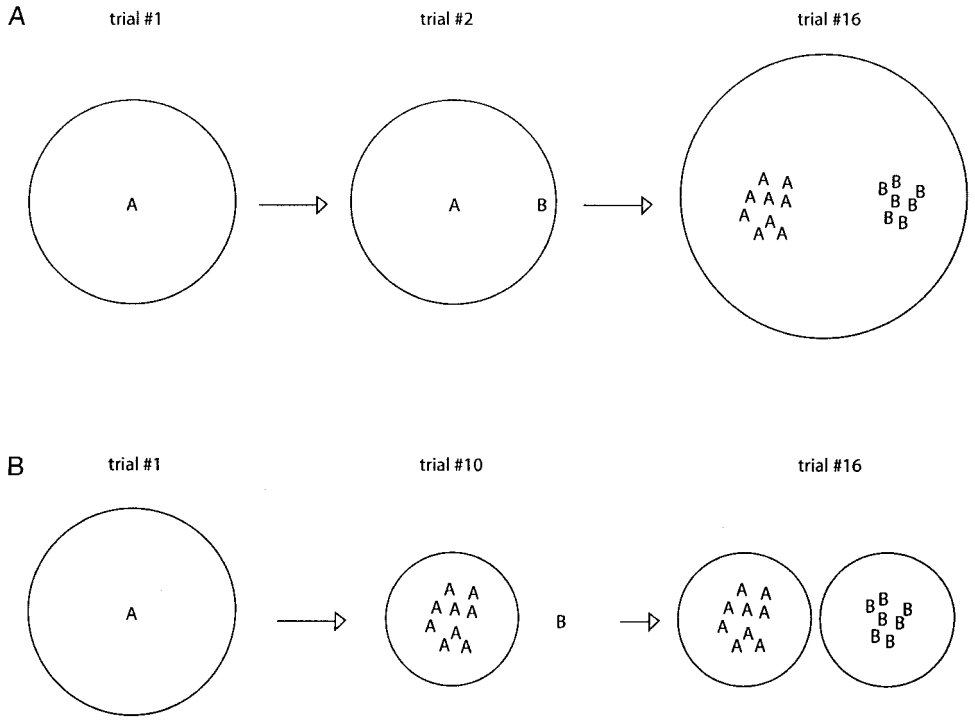

Figure 1. Illustrated predictions of an idealized category-learning model depending on whether patterns are shown in a mixed sequence (Panel A) or blocked/contrast sequence (Panel B).

amount of time they spend studying the correlated attributes in the list because these attributes become more predictable with learning (see Clapper & Bower, 1991; Heit, 1998; Son & Metcalfe, 2000; Stern, Marrs, Millar, & Cole, 1984); concomitantly, participants should increase the amount of time they spend studying the uncorrelated (unpredictable) attributes of each instance. Second, participants should progressively improve in memory both for the correlated attributes of each instance (because they have learned the patterns that make these attributes predictable) and for the uncorrelated attributes (because of the extra study time, participants increasingly allot to these unpredictable features).

Experiments 1 and 2 were intended to replicate the major pattern-sequence effects demonstrated earlier by Clapper and Bower (1994) by using this new study-time procedure rather than attribute listing. Demonstrating these effects with different tasks and stimulus materials would help establish their reliability and generality. To anticipate the results, in Experiments 1 and 2 we found practically no learning of the correlated features when instances of the two patterns were presented in random order without category labels. In contrast, in Experiment 3 we aimed to extend these earlier results by showing that participants could learn correlational patterns when presented with randomized instances, as long as cues (labels) were provided to help them discover that the stimuli could be sorted into separate categories. The General Discussion is devoted to formalizing our earlier intuitions about category-learning models by showing how specific assumptions adopted within J. R. Anderson's (1990, 1991) rational model of category learning permit that model to explain the sequence effects observed in our experiments.

Experiment 1

This experiment was a replication of the initial experiment of Clapper and Bower (1994) using the study-time procedure. Participants saw 12 *A* instances and 12 *B* instances; these were presented in a fully blocked order (all 12 *A*s, then 12 *B*s) for some participants and in a randomly mixed sequence for other participants. To compare the degree of unsupervised learning attained by the end of training in these two conditions, we structured the experiment so that all participants received a final sequence of four *A*s and four *B*s in a random mixed order. We expected the blocked sequence to produce far more learning than the mixed sequence. To evaluate whether participants who saw the mixed sequence displayed any learning at all, we compared their performance with that of a control group that was exposed to a series of randomly generated, "no-pattern" stimuli (i.e., no attributes were correlated across these instances).

Method

Participants. The participants were 43 undergraduate students of San Jose State University, who participated for partial fulfillment of their Introductory Psychology course requirement.

Materials. The training instances were verbal descriptions of 32 fictitious trees, presented in a column-list format. The instances were characterized in terms of 12 substitutive attributes, each with four possible values, yielding a stimulus set of 4^{12} possible instances. Examples of these attributes included the color of the tree's bark (dark gray, deep brown, mossy green, or light tan), its form or overall shape (low and shrub-like, tall and column-like, massive and wide branching, or twisting and vine-like), and the season in which it flowered (spring, summer, winter, or autumn).

Procedure. Each trial of the experiment consisted of a study phase followed by a test phase. At the beginning of the study phase, a verbal list was presented in the middle of a computer screen. At the top of the list was the name of a fictitious tree instance (these were arbitrarily selected Latin names from a plant identification guide), below which appeared a list of 12 verbal feature descriptors (attribute values). The attributes were presented one per row in the same vertical order (screen locations) on each trial, different for each participant (see Figure 2A). Every tree instance had a different name, so there was no suggestion that the name referred to a class or species.

At the start of the trial, each descriptor was masked by a row of *X*s (see Figure 2A). Starting with the cursor at a random position in the list, participants could look at the descriptors by pressing a designated *up* or *down* key, which unmasked and exposed the item (attribute value) on a given line. The exposed attribute value was covered up by *X*s again as soon as the participant moved the inspection pointer to a new line (attribute). Participants were limited to a total study time of 24 s for the entire 12-item list. The computer recorded the total amount of time a participant spent looking at each attribute of each instance.

In the test phase, participants were tested for their memory for the values of all 12 attributes of the just-studied instance. The 12 test questions were presented one at a time in random order in a multiple-choice format (see Figure 2B). The name of the most recent tree instance appeared at the top of the multiple-choice display, with four alternative answers (the four different values of an attribute) below. Participants tried to remember which of these values occurred in the last-studied instance and typed in the number (1–4) corresponding to that choice on their computer keyboard. (These tests were self-paced, though participants usually answered them within 3–5 s). Following this response, the computer displayed either a *correct* or an *incorrect* prompt under the test display, which remained on the screen. If the response was incorrect, that was indicated and the correct choice was indicated by an arrow in the display (see Figure 2C). The participant then pressed the *Return* key to see the next test question regarding the instance of the current trial.

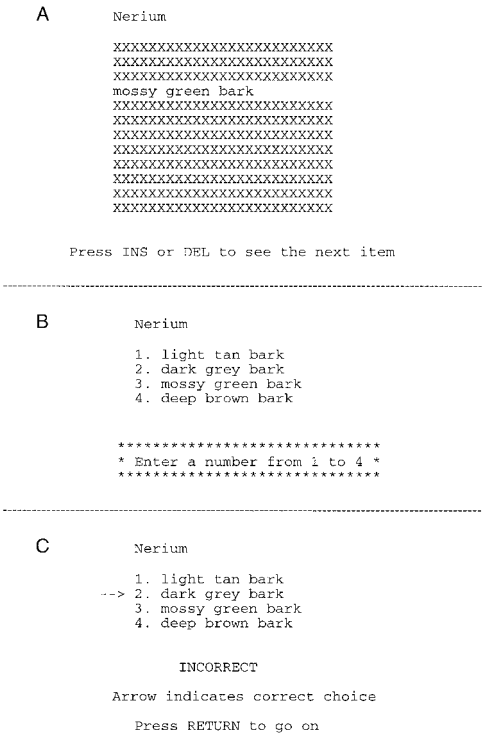

Figure 2. Computer display as it appeared during each phase (Panels A, B, and C, respectively) of a single trial from Experiments 1, 2, and 3. Each row in Panel A corresponds to an attribute of the instance.

After answering the 12 test questions about a given instance, participants received summary feedback regarding their memory for that instance. The percentage of items (out of 12) answered correctly on that trial was displayed and below this, the average percentage correct pooled over all test trials completed up to that point. If the trial score was higher than the cumulative score, the message "Good job! You beat your overall score!" appeared on the screen; if not, the message "Try to beat your overall score next trial" was displayed. If the participant answered all the test questions correctly on a given trial, the message "Good job! Your score was perfect!" was displayed. By pressing the *Return* key, participants moved on to studying the next instance in the training series.

The 12 attributes were tested in a different random order on each trial, and the order in which values were listed in the multiple-choice display was also randomized separately on each trial. The experiment consisted of a total of 32 such instance study–test trials.

Design. Participants were randomly assigned to three different conditions. In two of these conditions, nine attributes of the training stimuli had perfectly correlated values and three attributes were uncorrelated. Employing the notation of Table 1, the correlational patterns may be denoted as Pattern A = *111111111xxx* and Pattern B = *222222222xxx*, in which the first nine attributes have correlated values (denoted by *1*s or *2*s) and the last three attributes have uncorrelated values (denoted by *x*s). The uncorrelated attributes varied independently through all four values across different instances of a pattern or subset.

The two correlated conditions differed in the order in which the first 24 of 32 stimuli were presented. In the *blocked condition*, the first 12 stimuli were all examples of Pattern A and the second 12 stimuli were examples of Pattern B. The final eight trials provided a test series consisting of 4 *A*s and 4 *B*s presented in a randomly intermixed sequence. The *mixed condition* differed from the blocked condition only in the order in which the first 24 stimuli were presented, that is, the 12 *A*s and 12 *B*s were randomly intermixed for presentation rather than blocked by pattern. The randomization procedure was so constrained that no more than three stimuli from the same pattern appeared in a row. The final eight (test) trials for the mixed condition were identical to those of the blocked condition.

The uncorrelated or *control* participants saw a first set of 24 stimuli that had no correlational patterns among their attributes; all the attributes of these training stimuli varied independently. As in the correlated groups, 9 of the 12 attributes (serving as "pseudo" correlated attributes) varied through only two values in the training stimuli, and the remaining 3 attributes ("pseudo" uncorrelated attributes) varied through four values. Using notation analogous to that of Table 1, the stimuli in this condition may be denoted as *XXXXXXXXXxxx*, where the uppercase *X*s in the first nine positions denote attributes presented using two values and the lowercase *x*s in the last three positions denote attributes presented using all four values. (The former attributes were nonetheless always tested by asking that the correct value be selected from the four alternative values of those attributes.) The final eight (test) stimuli presented in the control condition were identical to those of the two correlated conditions, that is, they had nine correlated values as noted above. This final block is referred to as the *test block* for all three groups. Participants in the control condition were expected to show no learning before or during this test block, providing a baseline against which to evaluate any learning observed in the other two groups.

Counterbalancing. The stimuli for all participants in a given condition were generated by the testing program from the same input file, which contained coded specifications for generating the stimuli presented on each trial. Stimuli generated from these codes were presented in the same file order for all participants in all three conditions, allowing unbiased comparisons of learning across different groups. The correspondence between serial positions in the codes and the vertical position in which an attribute

was listed on the computer screen in the training instances was randomized for each participant. These random assignments were undertaken to balance out any idiosyncratic effects of particular attributes, values, or combinations of values, as well as vertical position in which the attributes were presented for study or test.

Results

The two dependent variables recorded on each trial of this experiment were (a) study times for correlated and uncorrelated attribute values during the study phase and (b) recognition-memory accuracy for correlated and uncorrelated values during the test phase. Because the total duration of the study period was a constant 24.00 s, any increase in study times to uncorrelated values would be reflected in a corresponding decrease in correlated value study times. Therefore, we describe the results in terms of the difference in average study times per attribute between uncorrelated and correlated values on a given trial, that is,

$$\frac{\sum t_{\text{uncorr}}}{3} - \frac{\sum t_{\text{corr}}}{9}.$$

The first term averages the times (t) over the three uncorrelated (uncorr) attributes; the second term averages times over the nine correlated (corr) attributes. We refer to these differences as *preference scores* because they reflect participants' average preference for studying uncorrelated rather than correlated values. The preference data for this experiment are shown in Figure 3.

Study times. In the blocked condition (see Figure 3A), the mean study time per attribute pooled over all 32 trials was 1.78 s for correlated values and 2.91 s for uncorrelated values, for an average difference of 1.13 s, $t(14) = 4.27$, $SE = 0.27$, $p < .01$. Turning to the mixed condition, no significant difference was observed between the overall study times for uncorrelated and correlated values (2.04 and 2.07 s, respectively), $t(14) = 0.60$, $SE = 0.06$, $p > .50$. For the control participants, the mean study times for the pseudo-correlated versus pseudo-uncorrelated values were nearly equal and were similar to those of the mixed condition. Comparing groups, we found that preference scores for the blocked condition were significantly greater than those in the mixed condition, $t(28) = 4.29$, $SE = 0.27$, $p < .01$; and the control condition, $t(26) = 3.63$, $SE = 0.30$, $p < .01$, overall; and also over the final eight-trial test block, $t(26) = 2.81$, $SE = 0.40$, $p < .01$, for blocked versus control, $t(28) = 2.78$, $SE = 0.37$, $p < .01$, for blocked versus mixed.

Figure 3A shows clear learning trends over trials within the blocked condition but not within the mixed or control conditions. For the blocked conditions, the per attribute preference for studying uncorrelated values increased throughout the *A* block, from 0.18 s on the 1st trial to 2.01 s on the 12th, linear trend $t(14) = 2.86$, $SE = 0.70$, $p < .02$. Preference scores for the blocked condition also rose rapidly during the *B* block, asymptoting by the 6th *B* instance, linear trend $t(14) = 4.04$, $SE = 0.25$, $p < .01$. Preference scores decreased slightly as the final test series began and remained somewhat depressed throughout the test block compared with the eight preceding *B* trials, $t(14) = 2.62$, $SE = 0.27$, $p < .05$. Nonetheless, Figure 3A shows clearly that preference remained positive throughout the test block. Excluding the first *A* instance, preference scores for *A* versus *B* instances did not differ reliably during the test block, $t(14) = 0.04$, $SE = 0.15$, $p > .50$.

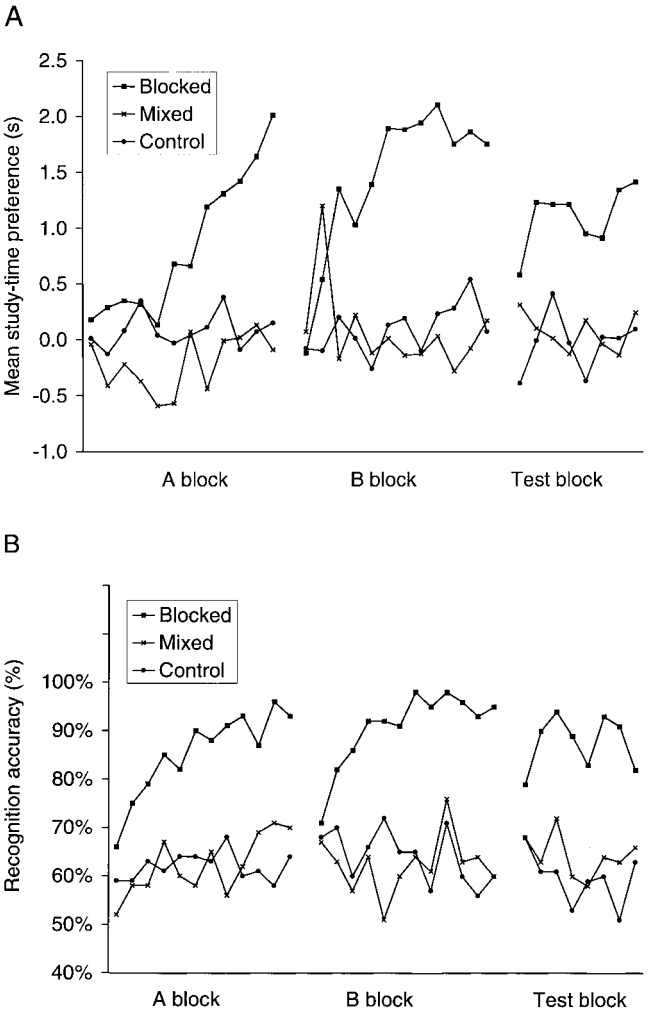

Figure 3. Study-time (Panel A) and recognition-memory accuracy data (Panel B) from Experiment 1. Trials are shown in their original order; the functions are disconnected to indicate where the *A* and *B* blocks are separated in the blocked condition and where the test block begins in all three conditions.

Recognition memory. The study-time results were strongly echoed by the recognition-memory data (see Figure 3B). Recognition accuracy overall was significantly greater in the blocked condition than in the mixed and control conditions, $t(28) = 7.38$, $SE = 0.03$, $p < .01$, and $t(26) = 7.07$, $SE = 0.04$, $p < .01$, respectively. Separating the memory data into correlated values versus uncorrelated values, accuracy in the blocked condition was greater than in the mixed condition for both types of attributes. This accuracy difference between conditions averaged 27.0% for correlated values, $t(28) = 7.86$, $SE = 0.03$, $p < .01$, and 23.0% for uncorrelated values, $t(28) = 5.99$, $SE = 0.04$, $p < .01$. None of the comparisons between mixed and control group data approached statistical significance.

In the blocked condition, recognition accuracy increased for both correlated and uncorrelated values over the first several trials

of both the *A* and *B* blocks. Averaged across correlated and uncorrelated values, accuracy in the blocked condition increased significantly over the first six instances of both patterns, $t(14) = 3.94$, $SE = 0.03$, $p < .01$ for Pattern A, $t(14) = 4.71$, $SE = 0.02$, $p < .01$ for Pattern B, but leveled out over the last six instances of each pattern. Their recognition accuracy during the test block was slightly below that of the preceding six trials, $t(14) = 3.30$, $SE = 0.02$, $p < .01$, and did not differ between Pattern A versus Pattern B, $t(14) = 1.07$, $SE = 0.02$, $p > .15$. No clear trends appeared in the memory data from the mixed or control condition.

Although the overall pattern of results over trials was similar for correlated values versus uncorrelated values in the blocked condition, memory for correlated values was greater overall (0.93 vs. 0.83), $t(14) = 5.45$, $SE = 0.02$, $p < .01$. This was also true in the mixed condition (0.65 vs. 0.60), $t(14) = 3.01$, $SE = 0.02$, $p < .01$, and in the control condition (0.66 vs. 0.59), $t(14) = 4.11$, $SE = 0.02$, $p < .01$. The greater accuracy for correlated attributes versus uncorrelated attributes in the latter two conditions was probably due to the fact that only two values of the correlated (or pseudo-correlated) attributes were ever presented in the study lists, whereas all four values of the uncorrelated attributes were shown throughout training (recall that four alternative values were shown for both types of attributes during each memory test). If some participants in the mixed and control conditions were sensitive to this difference, then their probability of guessing correctly on the recognition tests would have been somewhat greater for correlated than for uncorrelated attributes, as was observed.

Discussion

The results of this experiment replicated earlier ones (Clapper & Bower, 1994; Medin & Bettger, 1994) and extended their generality by using a different task, measure, and type of stimulus materials. When instances of the two patterns were presented in separate blocks, learning was rapid and remained evident during the final mixed test block. This superior performance during the final mixed test trials in the blocked condition indicates that it was not due merely to localized habituation to runs of repeated correlated values; rather, their mixed test performance suggests acquisition of two stable pattern representations that withstand randomized tests (see Clapper & Bower, 1991, 1994). By contrast, no significant learning was observed in the mixed condition of this experiment.

This blocking effect presents difficulties for several correlation tracking models that are insensitive to the order of seeing instances. For example, an optimal correlation learning algorithm (without memory limitations) that maintains and updates accurate correlational statistics on a given stimulus set would not be especially sensitive to order effects. This same insensitivity would be true of exemplar-storage models such as those of Medin and Shaffer (1978) and Hintzman (1986). Furthermore, many connectionist autoassociators actually predict stronger, more stable learning in mixed than in blocked sequences because such models would suffer from "catastrophic interference" in the blocked sequence. *Catastrophic interference* refers to the tendency of connectionist models to unlearn earlier patterns when the network weights are adjusted to learn the more recent patterns (e.g., Sharkey & Sharkey, 1995). Thus, the blocking effect eliminates at least some versions of correlation tracking, reducing the search space for valid models of unsupervised learning.

Experiment 2

Our goal for Experiment 2 was to replicate a second sequence effect demonstrated by Clapper and Bower (1994) by using the present list-study task rather than the attribute-listing task. That earlier experiment compared unsupervised learning of two groups of participants: One group received a long block of *A* instances before a test series of randomly mixed *A*s and *B*s; the second group received a mixed series, of which half were *A*s and the other half, *B*s, before proceeding to that final test series of more *A*s and *B*s. Those data showed strong learning of both categories by the participants in the former (contrast) condition, but minimal learning by participants in the latter (mixed) condition, even though eight more *B* instances had been presented in the mixed condition. As noted in the introduction, we were interested in replicating this contrast effect with the study-time measure because it presents a critical challenge for incremental correlation-tracking models.

Method

Participants, materials, and procedure. The participants were 31 students of San Jose State University, who participated for partial fulfillment of their Introductory Psychology course requirement. The individual trials, procedures, and materials were similar to those of Experiment 1. As before, the training instances were partitioned into the same two patterns based on correlations among 9 of the 12 stimulus attributes (Pattern A = *111111111xxx* and Pattern B = *222222222xxx*, where the *1*s and *2*s indicate correlated attribute values and the *x*s indicate uncorrelated attributes that vary independently through four possible values).

Design. Participants were randomly assigned to two conditions differing only in the sequencing of their training instances before a critical test series. In the *contrast condition*, only instances of Pattern A were presented for the first 16 trials, referred to as the *pretraining* block. In the second, the *mixed condition*, the pretraining block consisted of eight *A* instances and eight *B* instances presented mixed together in a random order. After this pretraining, the two groups then received the same random test series of 12 *A*s and 12 *B*s.

In both conditions, instances were constructed so that all four values of each uncorrelated attribute occurred an equal number of times within each category; within this constraint, values of these attributes were assigned randomly. The same stimulus set was presented to all participants in a given condition, but the order of instances within the pretraining and test blocks was randomized anew for each participant.

Results

The same types of data were collected in this experiment as in Experiment 1. The average study times and recognition-memory accuracies are displayed in Figure 4.

Study times. To begin with the study-time data (Figure 4A), in the contrast condition uncorrelated values were studied overall 1.33 s longer than were correlated values, $t(16) = 4.11$, $SE = 0.32$, $p < .01$. Uncorrelated values were also studied slightly longer than correlated values in the mixed condition, but this difference was not significant, $t(13) = 1.69$, $SE = 0.14$, $p > .10$. In direct comparisons, preference scores (uncorrelated study times minus correlated study times) in the contrast condition significantly exceeded those in the mixed condition, $t(29) = 2.89$,

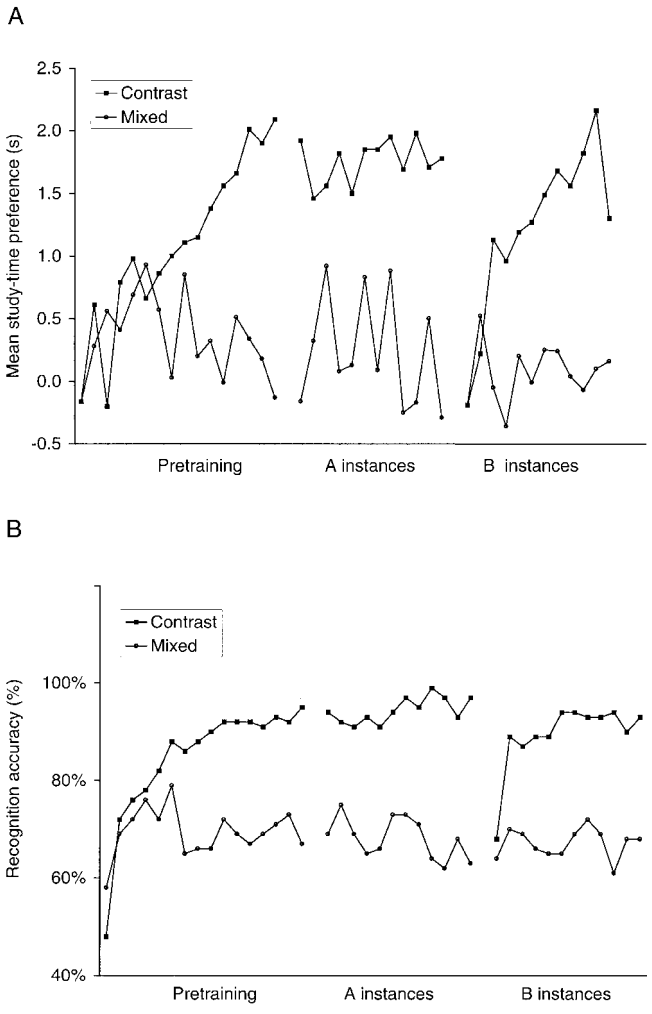

Figure 4. Study-time (Panel A) and recognition-memory accuracy data (Panel B) from Experiment 2. Pretraining trials are shown in their original order and separated from the test trials that follow. The randomized test trials are plotted separately by category in both conditions (the *A* instances before the *B* instances).

$SE = 0.38$, $p < .01$. Comparing only the final test block (identical for both conditions), we found that the effect remained significant, $t(29) = 3.01$, $SE = 0.43$, $p < .01$. The preference differences between the contrast and mixed conditions were also significant when Patterns A and B were analyzed separately, $t(29) = 2.99$, $SE = 0.49$, $p < .01$, for Pattern A; $t(29) = 2.94$, $SE = 0.39$, $p < .01$, for Pattern B.

Figure 4A shows strong learning trends over trials in the contrast condition but no significant trends in the mixed condition throughout its 16 pretraining or 24 test trials. Preference scores in the contrast condition showed a significantly increasing linear trend over Pattern A pretraining trials, $t(16) = 2.72$, $SE = 1.05$, $p < .02$. Preference scores for *B* instances for this group increased rapidly over the final 12 trials, $t(16) = 4.31$, $SE = 0.42$, $p < .01$, implying rapid learning of Pattern B. In fact, Pattern B learning in the contrast condition was so rapid that final preference scores for Patterns A and B (averaged over the last three instances of each) did not differ significantly, $t(16) = 0.35$, $SE = 0.19$, $p > .50$.

Recognition memory. Overall recognition accuracy (see Figure 4B) was significantly greater in the contrast condition than in the mixed condition throughout the pretraining and in the final test block: $t(29) = 7.51$, $SE = 0.03$, for Pattern A; $t(29) = 5.82$, $SE = 0.04$, for Pattern B; $p < .01$ for both. The contrast participants outperformed the mixed participants on memory accuracy for both Patterns A and B for both correlated values, $t(29) = 8.04$, $SE = 0.03$, $p < .01$, and uncorrelated values, $t(29) = 4.66$, $SE = 0.04$, $p < .01$.

As with the study-time measure, the recognition data revealed clear learning curves for each pattern in the contrast condition, whereas no significant trends appeared in the mixed condition. Overall recognition accuracy increased significantly over the first eight trials of the contrast pretraining block consisting of all *A*s, $t(16) = 8.35$, $SE = 0.03$, $p < .01$, and over the first eight B trials of the test series, $t(16) = 3.86$, $SE = 0.04$, $p < .01$, remaining stable thereafter. For contrast participants, asymptotic recognition accuracy for values of *B* and *A* instances did not differ significantly: averaged over the last six instances of each, $t(16) = 1.51$, $SE = 0.02$, $p > .10$.

Correlated values were remembered more accurately than uncorrelated values in both the contrast condition (0.94 vs. 0.84), $t(16) = 5.76$, $SE = 0.02$, $p < .01$, and the mixed condition (0.72 vs. 0.64), $t(13) = 3.19$, $SE = 0.02$, $p < .01$. Because the study-time data show little evidence of learning in the mixed condition, their greater accuracy in verifying correlated compared with uncorrelated values was probably due to the benefit of guessing the correct values of correlated attributes, which had only two values presented across the training instances, than of uncorrelated attributes, which had four values across instances.

Discussion

Experiment 2 results revealed rapid learning of both patterns in the contrast condition but little or no learning of either pattern in the mixed condition. These results essentially replicate those obtained by Clapper and Bower (1994). Theoretically, the most important feature of the present results is the demonstration that reducing (to zero) the number of training instances from one pattern (B) can sometimes increase later learning of that pattern. This contrast effect appears to be incompatible with simple correlation-tracking models. Incremental correlation tracking proceeds by strengthening measures of interfeature correlation as each relevant instance is encountered. In such theories, learning and strengthening of a pattern via associations or rules should increase monotonically with its exposure and practice. But our contrast condition removed all eight instances of Pattern B during pretraining and replaced them with more instances of Pattern A. Our finding that eliminating early *B* instances led to a major improvement in later *B* learning is incompatible with monotonic correlation tracking.[2]

[2] More precisely, we claim that our contrast effect disconfirms incremental, correlation-tracking algorithms that update measures of interfeature correlation (e.g., by testing rules or strengthening connections) following each presented training instance. Correlation-tracking models of unsupervised learning include rule-learning models such as those of Davis (1985) and Billman and Heit (1988), and simple autoassociators such as

We sum up our conclusions as follows: People in our experiments seemed to discover new patterns by suddenly noticing that an instance was very different from earlier instances about which they had developed some featural expectations. That noticing of sudden change triggered awareness of distinctly different subsets. However, if prior instances had been so variable as to undermine strong featural expectations (as with our mixed condition), then new, variable instances were seen as "more of the same" and were assimilated to the one category developed earlier. The intuition that unsupervised learning can be based on contrast detection is implicit in the logic of category invention. Category-invention models necessarily contain some mechanism for deciding whether each encountered stimulus is an acceptable member of some existing category, and if not, for generating a new category in response to such expectation failures (or perception of contrast with existing categories). Category-invention models do not necessarily expect learning of a target category to increase monotonically with the number of instances presented from that category. Invention models expect performance to improve rapidly with experience only after participants notice and identify valid subsets (categories) within a domain. But invention models expect little improvement if separate correlational patterns become lumped together into a single broad category, as happened with our mixed condition.

Experiment 3

The results of Experiments 1 and 2, along with the earlier ones of Clapper and Bower (1994), support our basic claims, namely: (a) that pattern-sequence effects on training are reliable and robust; (b) that these effects are compatible with models based on contrast detection, such as category-invention models, but not with exemplar storage or simple correlation-tracking models; and more generally, (c) that indirect methods such as attribute listing and the present study-time task provide workable procedures for investigating unsupervised learning.

In Experiment 3, we attempted to advance our argument for category invention by manipulating whether participants receive relevant category information (diagnostic labels) as they observe the randomized training stimuli. To understand the rationale of Experiment 3, begin by noting that participants in the mixed conditions of Experiments 1 and 2 showed little evidence of learning correlational patterns. Category-invention models explain this lack of learning as due to participants' aggregating *A* and *B* training instances into a single overgeneralized category (see Figure 1). From this perspective, if providing participants with diagnostic category labels were to prevent them from aggregating the earlier instances and instead induced them to segregate the stimuli into separate categories, then the patterns from mixed sequences should be learned just as they are from blocked and contrast sequences. We hypothesized that by presenting instances in a mixed sequence with diagnostic category labels, we could induce successful learning because participants would discover and use these labels as a basis for partitioning instances into different mental categories and learn the covariations within each.

By contrast, the correlation-tracking perspective might explain the lack of learning in our earlier mixed conditions (of Experiments 1 and 2) by arguing that participants were unable to accumulate correlational information because of a larger memory load, interference (Crowder, 1976; Postman 1971), or an excessively large search space (for rule-learning models, see e.g,. Billman & Heit, 1988; Davis, 1985). On this view, providing diagnostic labels for the training instances would be equivalent to adding just one more correlated feature to patterns A and B (J. R. Anderson, 1990, 1991). But this addition would, if anything, be expected to retard learning by creating even poorer memory or by giving rise to an even larger search space of potential rules. In any event, there is little reason within the exemplar storage or correlation-tracking frameworks to expect that providing such labels would result in significant improvements in unsupervised learning.

Method

Participants, materials, and procedure. The participants were 24 students of Humboldt State University, who participated in exchange for extra credit in several psychology classes. The same tree-list stimuli were used as in Experiments 1 and 2; the stimuli were again partitioned into two patterns based on correlations among 9 of 12 attributes (Pattern A = *111111111xxx* and Pattern B = *222222222xxx*). As before, the labels were 32 different Latin names arbitrarily selected from plant identification guides. The experiment consisted of instructions, 32 trials, and debriefing. The procedure was identical to that of the mixed conditions in Experiments 1 and 2 except participants (a) were given 36 s rather than 24 s to study each list and (b) were tested on their memory for the label as well as the 12 features of each instance. The instance label was tested first on each trial, prior to any of the features. Participants chose between two alternatives in the label tests and between four alternatives in the feature tests.

Design. Participants were randomly assigned to two conditions. In the *diagnostic-labels condition*, two labels that were to serve as class names (different for each participant) were randomly selected from the 32 labels. All instances of Pattern A received one of these class labels (presented at the top of the lists as shown in Figure 1), and all instances of Pattern B received the other. Thus, the labels were perfect predictors of category membership in this condition (although participants were never told this). In the *nondiagnostic-labels condition*, all 32 labels were used with a different label assigned to each instance. During the memory test phase of each trial, participants in the diagnostic condition chose between their two labels; participants in the nondiagnostic condition also chose between two possible labels, one of which was the correct label and the other the label of the instance shown on the previous trial. This helped to ensure that both test labels would be fairly equal in familiarity throughout the experiment, as were the labels in the diagnostic condition. Except for testing memory for instance labels on each trial, the within-trial events in this condition were identical to those of the mixed conditions of Experiments 1 and 2.

Results

The same types of data were collected as in Experiments 1 and 2. The average study-time preferences (for uncorrelated values

those of J. A. Anderson (1977). However, it is important to note that the contrast effect does not eliminate the entire spectrum of models that might use co-occurrence rules to represent predictive structure within a domain, only that particular subset that derives these rules (associations) entirely through some form of direct, incremental correlation tracking. For example, one could propose a nonmonotonic learning process that used surprise (novel features) to look for predictive structure within a domain but then represented this predictive structure in terms of feature co-occurrence rules rather than explicit categories. More research is needed to determine how such an approach would compare with our category-invention framework (and whether the two approaches really differ in any meaningful way).

minus correlated values) and recognition-memory data are shown in Figure 5. Because there was no principled difference between the structure of Pattern A versus Pattern B or the way in which they were presented in this experiment, the data for both have been averaged together in Figure 5 and the discussion to follow.[3]

Study times. Over all trials, participants in the diagnostic-labels condition devoted an average of 1.14 s longer to studying uncorrelated values than correlated values (see Figure 5A), $t(9) = 3.01$, $SE = 0.38$, $p < .02$. By contrast, study times in the nondiagnostic-labels condition did not differ between correlated values versus uncorrelated values, $t(11) = 0.61$, $SE = 0.18$, $p > .50$. Thus, overall preference scores were significantly greater in the diagnostic condition than in the nondiagnostic condition, $t(18) = 2.61$, $SE = 0.41$, $p < .02$. This difference in study-time preference between the two conditions increased over trials: Although it was not statistically significant over the first eight trials, $t(18) = 1.53$, $SE = 0.32$, $p > .10$, it became so over the last eight, $t(18) = 2.56$, $SE = 0.63$, $p < .05$.

Linear trend analyses showed clear evidence of unsupervised learning in the diagnostic condition, with preference scores increasing significantly over trials, $t(11) = 2.83$, $SE = 1.45$, $p < .02$. Although the preference for studying uncorrelated values compared with correlated values was not significant overall in the nondiagnostic condition, a linear trend analysis did reveal a significant increase over the first eight trials, $t(11) = 3.67$, $SE = 0.22$, $p < .01$, but none over the last eight trials, $t(11) = 0.46$, $SE = 0.31$, $p > .50$. This pattern contrasts with that in the diagnostic-labels condition, in which an increase was observed over both early and late trials.

Recognition memory. The memory data also show greater unsupervised learning in the diagnostic-labels condition. Memory for correlated values (see Figure 5B) averaged over trials was significantly greater in the diagnostic than the nondiagnostic condition, $t(18) = 2.15$, $SE = 0.04$, $p < .05$. This difference was apparent by the third or fourth instance of a given category (see Figure 5B). However, the two conditions did not differ significantly in memory for uncorrelated features, as shown in Figure 5C, $t(22) = -0.86$, $SE = 0.05$, $p > .40$.

Recognition accuracy improved over trials in both conditions (see Figure 5B and 5C). In the diagnostic condition, memory for both correlated and uncorrelated values increased significantly over the first eight trials—$t(11) = 6.51$, $SE = 0.03$, $p < .01$, for correlated values, and $t(11) = 3.83$, $SE = 0.33$, $p < .01$, for uncorrelated values—but not over the last eight trials ($p > .50$ for both). In the nondiagnostic condition, memory for correlated values increased over the first eight trials, $t(11) = 5.58$, $SE = 0.02$, $p < .01$, but not over the last eight, $t(11) = 0.08$, $SE = 0.04$, $p > .50$, and memory for uncorrelated values did not increase throughout the experiment, $t(11) = 0.93$, $SE = 0.04$, $p > .25$, for the first eight trials, and $t(11) = -1.50$, $SE = 0.05$, $p > .15$ for the last eight trials.

Correlated values were remembered an average of 18.6% better than uncorrelated values in the diagnostic condition, $t(11) = 4.71$, $SE = 0.14$, $p < .01$, but only 6.1% better than uncorrelated values in the nondiagnostic condition, $t(11) = 4.12$, $SE = 0.01$, $p < .01$. As suggested for previous experiments, the small advantage for correlated values in the nondiagnostic condition was probably due to some of these participants learning that only two values were ever presented in the study lists for correlated attributes, compared with four values for uncorrelated attributes, resulting in a higher probability of correct guesses for correlated attributes.

Discussion

As expected, significant unsupervised learning occurred in the diagnostic-labels condition but relatively little in the nondiagnostic-labels condition in this experiment. Study time preferences increased more over trials in the diagnostic condition, significantly exceeding preference scores in the nondiagnostic condition over the second half of the experiment. Recognition accuracy for correlated features increased more in the diagnostic condition, significantly exceeding the increase in the nondiagnostic condition. A third prediction, that the increased study time allocated to uncorrelated values would improve their memory more in the diagnostic than in the nondiagnostic condition, was not confirmed in the present experiment. Given the facts that memory improvement accompanied unsupervised learning in Experiment 1 and 2 and that some indications of improvement over trials seemed to appear in the present data, this insignificant difference may simply be a statistical anomaly. Overall, the results of this experiment are consistent with the category-invention prediction that people can readily learn correlational patterns from mixed sequences when cues are provided that serve as category labels.

J. R. Anderson (1990, 1991) has pointed out that a category label for an instance is logically no different from other correlated features such as its color or size. From this view, the labels in the diagnostic-labels condition should have been treated no differently than any of the other correlated features of the instances; all were equally diagnostic of the current pattern and predictive of each other. For example, participants might have classified the training instances equally well in terms of brown bark versus gray bark; if these were part of the correlated values, this classification would have been as valid as a classification in terms of the category labels.

Despite this logical equivalence of category labels to other correlated features, the labels clearly had a disproportionate influence on our participants' ability to learn the predictive structure of the stimulus set. (Compare this with the mixed conditions of Experiments 1 and 2.) Moreover, the influence of the labels cannot be ascribed to a simple, incremental effect of adding one more correlated feature to the training stimuli. As Bloom (2000) has emphasized (see also Yamauchi & Markman, 1998), labels are treated as having special communicative status and intent in language, providing them an extra degree of salience. Because of the prominence and task relevance of these labels (e.g., they appeared in a salient location at the top of each list and obviously referred

[3] We eliminated a small number of outlier participants in this experiment whose memory and/or study times deviated substantially from their group norm. Thus, some of the reported analyses are based on trimmed data, in which the highest and lowest scores from each group were eliminated and comparisons were then made using the remaining, nonoutlier scores. This procedure is fair, does not bias the outcome in any way, and provides a clearer picture of our experimental results. The reader can identify these analyses based on the lower degrees of freedom assumed for the within- and between-groups analyses (for untrimmed data these degrees of freedom are 11 and 22, respectively, whereas for trimmed data they are 9 and 18).

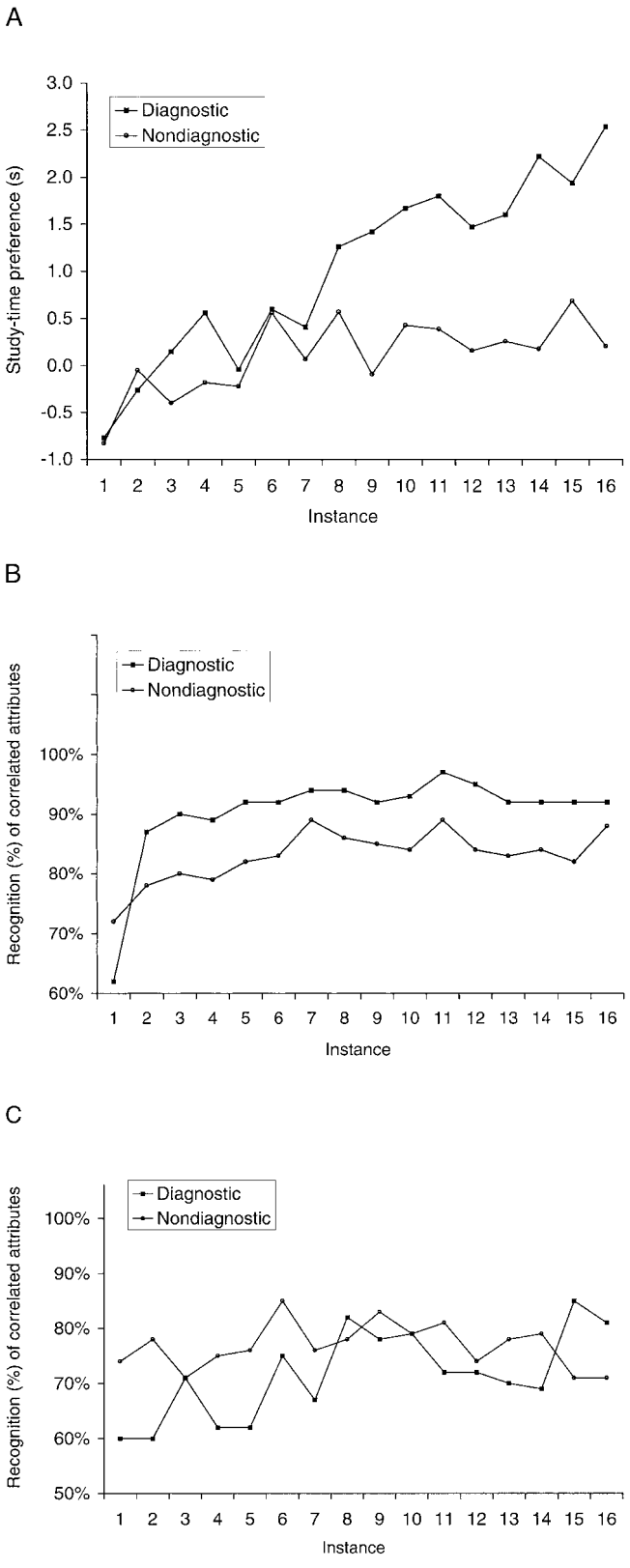

Figure 5. Study time (Panel A) and recognition-memory accuracy data (Panels B and C) from Experiment 3. Trials are shown averaged over *A* and *B* categories. Recognition data for correlated attributes are shown in Panel B and those for uncorrelated attributes are shown in Panel C.

to taxonomic categories), our participants apparently used them as a basis for partitioning the stimulus set into two separate categories. (In theory, a similar advantage could have been conferred to any of the correlated values by framing it as a class label, as when we might first say about each instance, "This is a flowering [or nonflowering] type of tree.") This partitioning in turn enabled participants to learn the correlational structure within each subset.

It could be argued that our diagnostic-labels condition was a kind of supervised learning rather than unsupervised learning because categories were provided in advance by the experimenter in the form of diagnostic labels. However, although participants in the diagnostic condition had to remember the category label of each instance for the recognition-memory tests, they were never asked to guess which category an instance belonged to given only its features. In other words, the present task was not a categorization task, at least not directly. Further, there were no explicit rewards in the list-study task for learning which features were associated with a given category, only indirect payoffs such as improved performance on the instance memory task. Overall, participants had the option of ignoring category labels in the list-study task in a manner that would be impossible in a standard supervised learning task. Thus, the discovery problem, which we propose as the most interesting aspect of unsupervised learning, was still present for participants in the diagnostic-labels condition; the solution was simply far easier to find because of the availability of a more salient cue (i.e., the diagnostic labels) to the correlational structure of the stimulus set.

General Discussion

Taken together, the results of these experiments make a strong case for rejecting our null hypothesis that participants were acquiring detailed knowledge of the correlational structure of our stimulus sets merely by accumulating information about feature correlations across successive training instances. Both sequence and labeling effects suggest that learning in the study-time task depends crucially on noticing surprising changes in the stimulus patterns and creating new categories to accommodate these changes. After the new category is postulated, participants begin to accumulate regularity information conditional upon the category of instances.

A Rational Analysis of Unsupervised Learning

To formalize our intuitive characterization of how category learning could accommodate our results, we considered how an optimal or ideal category learner would perform in our experiments. J. R. Anderson's (1990, 1991) rational model provides a description of normative or optimal categorization on the basis of the assumption that the goal of categorization is to capture predictive structure in the environment, that is, to maximize the ability to predict the features of objects (instances) given partial information about them. The performance of participants in our experiments can be compared with this idealized model to shed light on how well our participants learned the patterns and on what their biases and performance limitations were.

The rational model begins by assuming that any presented object will be assigned to the category that is most probable given the feature structure of that object. A new category will be created for this object if its features are sufficiently improbable given any

of the existing categories. The probability of assigning an object to a given category is calculated as a Bayesian posterior probability:

$$P_k = P(k|\mathbf{F}) = \frac{P(k)P(\mathbf{F}|k)}{\sum_k [P(k)P(\mathbf{F}|k)]}, \quad (1)$$

where k is the target category, $\mathbf{F}$ is the feature vector (description) of the current object, $P(k|\mathbf{F})$ is the probability of the target category given the feature vector of the current instance, $P(k)$ is the prior probability of category k, and $P(\mathbf{F}|k)$ is the probability of the feature vector of the current instance given category k. Thus, for each category k, the model must compute the prior probability of the category, $P(k)$, and the conditional probability of the instance given that category, $P(\mathbf{F}|k)$.

The details of how the rational model computes $P(k)$ and $P(\mathbf{F}|k)$ are provided in the Appendix. For present purposes, the important point is that both calculations depend on theoretical parameters that can be chosen to allow the rational model to simulate different assumptions about learners' biases, prior beliefs, and their limited memory. Thus, the estimated prior probability of any given category, $P(k)$, depends in part on a "coupling probability" parameter (denoted c), defined as the subjective likelihood that two randomly sampled objects from the stimulus set will be members of the same category (i.e., can be "coupled together"). Intuitively, a high value of this coupling parameter would correspond to the learner assuming in advance that only one or a few distinct categories exist within the entire set, thus decreasing the prior probability of a new category, P(new); in contrast, a low value of c would correspond to a prior belief that many categories are likely to exist within the set, thus increasing P(new) (see Appendix for details). Typical simulation values for this parameter when fitting experimental results range from .3 to .5 (e.g., J. R. Anderson, 1990, 1991; J. R. Anderson & Fincham, 1996).

The conditional probability of the instance's features given a particular category, $P(\mathbf{F}|k)$, is computed as a product of the individual probabilities of each value j of all attributes i of the instance, conditional upon its being a member of category k (see Appendix). These $P_i(j|k)$s depend on the learner's prior beliefs about the likely distribution of values on each attribute, combined with their actual observations of these attribute values over previous instances of the category. As shown in the Appendix, these feature-to-category associations may also depend on the learner's memory for observations (denoted δ): When memory is poor, observations have less impact relative to the learners' prior beliefs. In terms of Figure 1, we can imagine that the poorer the learner's memory, the less impact each presented instance would have on modifying category boundaries.

An important point to note about the rational model is that for most parameter values its performance is not at all like our participants'. Rather, the model usually assigns A and B instances to separate categories regardless of the sequencing of training instances. Thus, for its usual parameter settings, the model would show rapid learning and no significant differences between our mixed sequences versus blocked or contrast sequences. But all is not lost. To simulate our results within the rational model, we searched for parameter values (of c and δ) for which the model would aggregate the first B instance into the same category as one or more prior A instance(s) when these are presented in a mixed order. Table 2 illustrates how these parameters determine whether or not the model will create a new category on the first-B trial (preceded by a single A trial) in a typical mixed-sequence condition. This table shows that for sufficiently high values of c and low values of δ, the rational model will fail to create a new category on the first B trial after one A. Thus, the rational model can predict aggregation of different correlational patterns, but only if it assumes that participants' memory is poor and that the prior probability of new categories is low. (However, note how much these parameters contrast with those typically used by J. R. Anderson, 1990, 1991—in which $\delta = 1$ and c is low—when fitting supervised category-learning results.)

Given that parameter values are selected from within the range highlighted in Table 2, the rational model can predict patterns of learning that approximate those of our participants'. Figure 6 shows the learning predicted by the rational model (for $c = .90$ and $\delta = .40$) for our blocked, contrast, and mixed conditions.[4] Note that in the mixed conditions, the learning index $P_i(j|k)$ changes relatively little over trials (see Figures 6A and 6B), because the model aggregates the first B instance to the initial A category on trial 2 and this aggregation persists blindly throughout training. Because the patterns are both assigned to a single category and because this version of the rational model does not directly accumulate information about feature correlations within a category (but see J. R. Anderson & Fincham, 1996; J. R. Anderson & Matessa, 1992, for a version that does), the model is unable to acquire any knowledge about the correlational structure within the stimulus sets in this condition.[5]

The model behaves quite differently in the Blocked and Contrast conditions (illustrated in Figures 6A and 6B, respectively). In both conditions, the model increases its subjective probabilities (or learning index) $P_i(j|k)$ for the A-correlated values on each trial of the first (A) block. Despite the burden of poor memory and a bias against new categories, after several A trials these values are sufficiently probable within the A category that the first instance of B provides a very poor match, so that a new category is likely to be generated. Once a separate category is established to accommodate the first B instance, all subsequent instances are correctly assigned to A or B categories regardless of the sequence in which they occur. This assignment enables the model to learn the feature probabilities within each category.

[4] The index of learning shown in Figure 6 is the average conditional probability, $P_i(j|k)$, for the correlated values of each training instance. This index was chosen because increased recognition accuracies and decreased study times for correlated values were the primary index of learning in our experiments. These behavioral measures were assumed to depend directly on their conditional probabilities within the category to which each training instance was assigned.

[5] The slight increase in $P_i(j|k)$ that occurs over trials in the mixed condition is due to the model learning that the two presented values are the only ones ever to occur in instances of the aggregated category. In our simulation runs, the prior distributions for each attribute were set up with four equally probable values. Over trials, the subjective probability of the two presented values gradually increased at the expense of the two nonpresented values, resulting in the apparent learning trend displayed in Figure 6. The same slight increase for two-valued attributes would also occur in an uncorrelated control condition such as that used in Experiment 1 (assuming that the simulation program was initially set up to expect more than two possible values per attribute).

Table 2
Probability of a New Category on Trial 2 of an AB . . . Mixed Sequence, According to the Rational Model

	Memory parameter (δ)									
Coupling probability (c)	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
.1	.92	.94	.96	.97	.97	.98	.98	.99	.99	.99
.2	.84	.88	.91	.93	.94	.96	.97	.97	.98	.98
.3	.76	.81	.85	.88	.91	.93	.94	.95	.96	.97
.4	.67	.73	.78	.82	.86	.89	.91	.93	.94	.96
.5	.57	.64	.70	.76	.80	.84	.87	.90	.92	.94
.6	**.47**	.54	.61	.68	.73	.78	.82	.86	.88	.91
.7	**.37**	**.43**	.51	.57	.64	.70	.75	.79	.83	.86
.8	**.25**	**.31**	**.37**	**.44**	.51	.57	.63	.69	.74	.78
.9	**.13**	**.17**	**.21**	**.26**	**.31**	**.37**	**.43**	.50	.56	.62

Note. The rational model creates a new *B* category on trial 2 only if its estimated probability exceeds that of the existing *A* category, that is, when $P(\text{new}) > .50$. Hence, the boldface entries indicate parameter values for which the model aggregates Patterns A and B into a single category.

In summary, the rational model can provide an explanation for the sequence effects in our data—in particular, the nonmonotonicity implied by the contrast effect and the surprisingly poor learning in the mixed conditions despite highly separable categories. This analysis suggests that even an optimal category learner may show little accumulative learning from experience if it begins with prior beliefs that new categories are highly unlikely and must operate under severe constraints on its short-term memory. Thus, the poor learning in our mixed sequences that initially appeared highly suboptimal and counterintuitive can be shown to follow from a normative model of adaptive categorization, given reasonable assumptions about memory constraints and task-related biases.

To account for the labeling effects in Experiment 3, the rational model might assume that categories may be either (a) inferred from the stimulus data, as currently done, or (b) determined in advance by external factors such as hints regarding the experimenter's communicative intention or participants' strategies. In the latter case, the Bayesian analysis would proceed by treating the imposed categorization scheme (from the labels) as a given and then computing conditional feature probabilities based on this segregation. Thus, participants would use these labels as an external basis for categorizing the training instances, essentially classifying stimuli by rule rather than by overall similarity, as is usually assumed by the rational model.[6]

The parameter values we have had to assume to account for the present data are quite different from those used by J. R. Anderson (1990, 1991) to fit the results of typical category-learning tasks. However, these unusual parameter values seem appropriate for our experiments because of the complexity of the verbal descriptions of our instances (resulting in unusually poor memory, i.e., low δ) and the fact that participants acquired categories incidentally, without presetting their search behavior (i.e., they were never told to look for categories or given any reason to expect them to exist within our stimulus sets, which would be reflected in a high value for the coupling probability c). Presumably memory performance, and hence category detection, would be better if shorter description lists or pictorial stimuli were used. Indeed, the results of Clapper and Bower (1994), in which pictorial stimuli were used, did show somewhat more evidence of learning in the mixed condition than did the present experiments. Mixed-condition performance should also be better under intentional learning conditions (e.g., standard sorting or supervised classification tasks), in which participants expect separate categories and set an explicit goal to search for them (thus setting a lower value for c).

The present data, in combination with the rational analysis, strongly support our suggestion that participants invented discrete categories to learn the correlational patterns in these stimulus sets. At the same time, this analysis helps narrow down the class of psychologically valid category-learning models by specifying some of the features that such models must possess in order to accommodate our data. Obviously, nonincremental clustering models that simultaneously analyze large batches of training instances and compute optimal categories afterwards (e.g., Fried & Holyoak, 1984; Sattath & Tversky, 1977; Caroll, 1976; Shepard, 1962, 1980; Torgenson, 1952) cannot accommodate the sequence effects demonstrated here. More surprisingly, many incremental categorization models are eliminated as well, primarily because they do not take explicit account of learners' biases, prior beliefs, and/or memory limitations (e.g., Ahn & Medin, 1992; Feigenbaum 1963; Fisher & Langley, 1990; Kolodner, 1983; Lebowitz 1987; and many others). Most of these models are designed to extract categories from the training data as efficiently as possible. They predict that our patterns would be trivially easy to learn regardless of their order of presentation, contrary to fact.

An advantage of the Bayesian framework is that it helps us to understand the relatively poor learning in our mixed conditions by postulating variable parameter values that can be related intuitively to memory limitations and prior task biases. Although it seems

[6] Another possibility, suggested by J. R. Anderson (1990, 1991) is that the presence of category labels may cause learners to decrease the weight of their feature prior probabilities α_j and α_k relative to observed features, thus improving category learning. However, it seems likely that participants in our experiments were forced to assign prior probabilities a high weight relative to observations simply because their memory for observations was so poor that they had no other choice. Such basic memory limitations would seem to preclude a solution in which one simply reweighs one's priors so that earlier observations may have a greater impact.

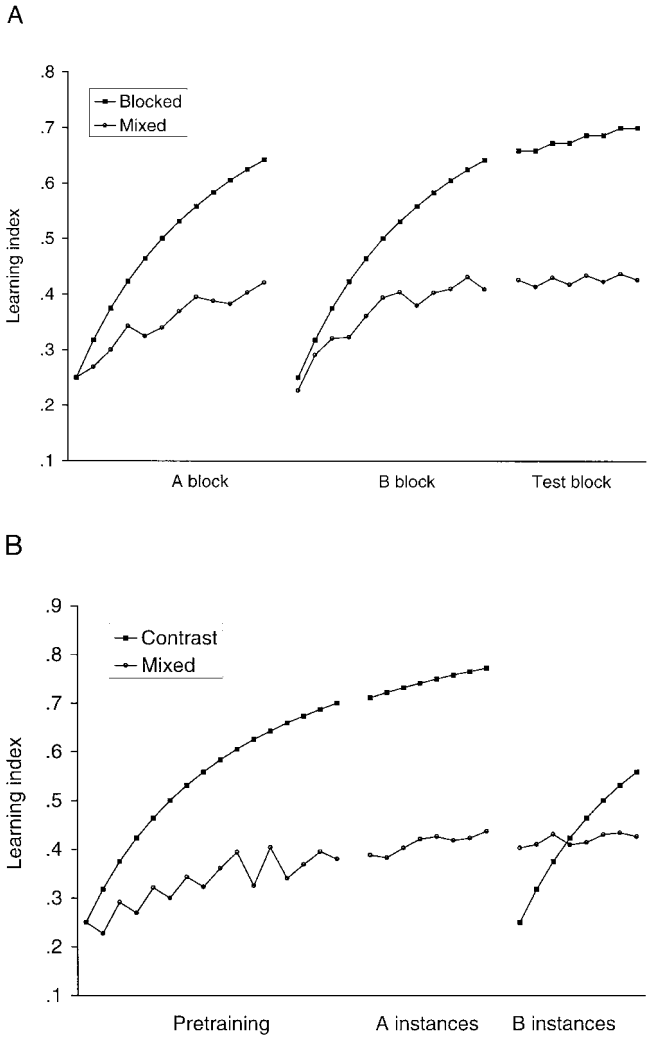

Figure 6. Predictions of the rational model of Anderson (1990, 1991) for different training sequences. The learning index is the probability of the correlated values of the current instance within the assigned category, which can be compared with the learning data from our experiments.

unlikely that the simple rational model presented here could provide a complete or fully adequate account of human categorization (e.g., see the critique by Murphy, 1993), it does provide a coherent framework within which many factors can be described and their effects predicted. It remains to be seen whether other approaches can be modified to provide equally compelling accounts of our data.

Implications for Real-World Learning

What does this research imply about people's learning in the everyday world? As in our experiments, people often seem fully engaged with whatever goal they are pursuing at the moment and appear to have no explicit goal of inventing categories and computing generalizations. Moreover, human memory for the fine details of experience is rather poor and especially so for features attended to only briefly in passing (as is true for much of what we encounter in daily life). Thus, the distinctive characteristics of our task—prior biases against searching for new categories and relatively poor memory for complex individual cases—may be representative of many informal discovery-learning situations that people face every day.

An interesting possibility suggested by the present results is that people may remain essentially "blind" to a substantial amount of the predictive structure in their environments, despite extensive experience. For example, most people have a sense that things like trees, songbirds, crockery, electronic equipment, architectural styles, and so on vary along many dimensions and are probably separable into reasonably distinct categories beyond those they attend to. Although in many cases informal experience seems to teach people the dimensions of variation within such domains, they still may fail to learn the correlational structure that defines distinct categories. For example, one might observe that shorebirds differ in many ways, but without determined effort or outside assistance, one may never acquire a clear sense of the individual species and their characteristics. We suggest that such failures of unsupervised discovery may be due to aggregation processes like those that retarded learning in our mixed conditions. We are not suggesting here that people are unable to discover for themselves predictive structure in such domains; rather, we point out (and explain why) mere exposure to many examples may not guarantee that observers will automatically and effortlessly discover all the predictive structure that exists in a given domain. It often takes special efforts to learn the differentiating refinements that permit subdivision of a general class into more highly informative subordinate classes.

Such shortcomings in human learning might seem paradoxical from an adaptionist or evolutionary perspective (e.g., Barkow, Cosmides, & Tooby, 1992; Buss, 1999), but we suspect that they have little impact on people's actual ability to cope with the demands of their day-to-day lives. Major distinctions among basic objects in the environment (e.g., trees vs. grasses vs. birds vs. rocks) typically arise from differences in the presence versus the absence of basic stimulus attributes rather than their specific values. These glaring differences render virtually impossible a learner's aggregating them regardless of their initial biases and memory limitations. Moreover, our analysis suggests that people can probably detect predictive structure through deliberate search (setting c low and δ relatively high) that would otherwise be overlooked in the absence of such explicit goals and expectations. Although people's failure to discover distinct categories of pine trees, grasses, and shorebirds may carry little adaptive cost in their daily lives, it is our hope that exploring such limitations in the laboratory may prove fruitful for understanding the cognitive machinery underlying human learning.

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 27–90). Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 259–277.

Anderson, J. R., & Matessa, M. (1992). Explorations of an incremental, Baysian algorithm for categorization. *Machine Learning, 9,* 275–308.

Barkow, J., Cosmides, L., & Tooby, J. (1992). *The adapted mind.* New York: Oxford University Press.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science, 12,* 587–625.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 458–475.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Buss, D. M. (1999). *Evolutionary psychology: The new science of the mind.* Needham Heights, MA: Allyn & Bacon.

Caroll, J. D. (1976). Spatial, non-spatial, and hybrid models for scaling. *Psychometrika, 41,* 439–463.

Clapper, J. P., & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 27, pp. 65–108). New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 443–460.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Davis, B. R. (1985). An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15,* 570–579.

Feigenbaum, E. A. (1963). A simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 297–309). New York: McGraw-Hill.

Fisher, D., & Langley, P. (1990). The structure and formation of natural categories. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 241–284). San Diego, CA: Academic Press.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Gluck, M., & Corter, J. (1985). Information, uncertainty, and the utility of categories. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society* (pp. 283–287). Hillsdale, NJ: Erlbaum.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 712–731.

Hintzman, D. L. (1986). "Schema abstraction" in a multi-trace memory model. *Psychological Review, 93,* 411–428.

Homa, D., & Cultice, J. (1984). The role of feedback, category size, and stimulus distortion in the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 83–94.

Kaplan, A. S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition, 27,* 699–712.

Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science, 7,* 281–328.

Lebowitz, A. M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning, 2,* 103–138.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and categorically related examples. *Psychonomic Bulletin & Review, 1,* 250–254.

Medin, D. L., & Shaffer, M. M. (1978). A context theory of classification learning. *Psychological Review,* 85, 207–238.

Murphy, G. L. (1993). A rational theory of concepts. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 29, pp. 327–339). New York: Academic Press.

Postman, L. (1971). Transfer, interference, and forgetting. In J. W. Kling & L. A. Riggs (Eds.), *Experimental psychology* (3rd ed., pp. 1019–1132). New York: Holt, Rinehart & Winston.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 45–77). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery in competitive learning. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 151–193). Cambridge, MA: MIT Press.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42,* 319–345.

Sharkey, N. E., & Sharkey, A. J. C. (1995). An analysis of catastrophic interference. *Cognitive Science, 7,* 301–329.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1980, October 24). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 290–298.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 204–221.

Stern, L. D., Marrs, S., Millar, M. G., & Cole, E. (1984). Processing time and the recall of inconsistent and consistent behaviors of individuals and groups. *Journal of Personality and Social Psychology, 47,* 253–262.

Torgenson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrica, 30,* 379–393.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124–148.

Zeaman, D., & House, B. J. (1963). The role of attention in retardate discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency* (pp. 159–223). New York: McGraw-Hill.

Appendix

Description of Relational Model

The rational model (see J. R. Anderson 1990, 1991) attempts to compute the prior probability *P(k)* of each category *k*, and the conditional probability of the instance's feature vector **F** given each category *k*, $P(\mathbf{F}|k)$. The procedure for computing *P(k)* depends on whether *k* is an already-existing category or a new category. For existing categories, this prior probability is computed according to the following formula:

$$P(k) = \frac{cn_k}{(1 - c) + cn}, \quad \text{(A1)}$$

where n_k is the number of previous instances (objects) assigned to category *k*, *n* is the total number of objects seen so far, and *c* is the coupling probability. The probability that the object comes from a new category is one minus the sum of the *P(k)*s, which comes to

$$P(\text{new}) = \frac{(1 - c)}{(1 - c) + cn}. \quad \text{(A2)}$$

The second element in Equation 1 (presented in the General Discussion), the conditional probability of the current instance given category *k*, is calculated as:

$$P(\mathbf{F}|k) = \prod_i P_i(j|k), \quad \text{(A3)}$$

where values *j* on attribute *i* make up the feature set **F**. Here, $P_i(j|k)$ refers to the probability that a member of category *k* has value *j* on attribute *i*. The individual $P_i(j|k)$s are computed by combining observations over previous instances of category *k* with prior expectations concerning the probability of value *j* of attribute *i*, as follows:

$$P_i(j|k) = \frac{n_j + \alpha_j}{n_k + \alpha_k}. \quad \text{(A4)}$$

In Equation A4, n_k is the number of previous objects in category *k*, n_j is the number of objects in category *k* with the same value *j* on dimension *i* as the current object, α_j represents the strength of participants' prior belief that value *j* will occur, and α_k represents the sum of the α_j for all values *j* of attribute *i*. This equation can be viewed as a weighted combination of the empirical proportion n_j/n_k (objects in category *k* with value *j* of attribute *i*) and learners' prior probability of value *j* in category *k*, α_j/α_k (J. R. Anderson 1990, 1991).

It is important to note that the rational model assumes feature independence within categories, that is, individual $P_i(j|k)$s are weighted equally and then multiplied to produce an overall estimate of the conditional probability of the instance given the current category. The model captures nonindependence (correlational structure) among features within a set by creating subsets (categories) within which attributes are assumed to vary independently (J. R. Anderson, 1990, 1991; J. R. Anderson & Fincham, 1996; J. R. Anderson & Matessa, 1992). Regarding our experiments, it may seem odd to consider features that occurred together with 100% probability within a category to be "independent." However, the feature cooccurrences in our stimulus sets were completely predictable on the basis of category membership—essentially, they were a mere side effect of the fact that certain features were perfectly correlated with certain categories and, incidentally, with other features that also happened to be correlated with the same categories. (In fact, any feature that occurs with 100% probability must be considered independent of all other features within a category simply because its presence does not depend on those other features in any way—it is guaranteed by category membership.) As for the uncorrelated attributes, our stimulus sets were specifically constructed so that these attributes would vary independently of each other and the categories.

In J. R. Anderson's (1990; 1991) original formulation, memory for training instances is assumed to be perfect, that is, the optimal model includes no explicit parameter to reflect limited memory or forgetting. One way to express the impact of limited memory is to add a memory parameter δ to Equation A4 to reflect participants' overall level of certainty regarding their instance memory, namely,

$$P_i(j|k) = \frac{\delta n_j + \alpha_j}{\delta n_k + \alpha_k}. \quad \text{(A5)}$$

The memory parameter δ ranges from 0.00 (full forgetting) to 1.00 (complete memory) and weighs participants' observations of past instances by how much they forget. Low confidence concerning past observations (instances) could also be captured indirectly by setting δ to 1.00 and increasing the values α_j and α_k in Equation A4 while leaving the ratio of α_j to α_k constant. Increasing these values weighs the participants' priors more heavily relative to observed instances in computing the $P(j|k)$s; this would decrease the impact of empirical observations and mimic the effects of limited memory. Although adding the δ parameter seems a more intuitive way to characterize forgetting than does altering the weight of prior beliefs, it does have the disadvantage of adding a new parameter to J. R. Anderson's (1990) original model.

Received February 6, 2001
Revision received March 15, 2002
Accepted March 26, 2002 ■